Nucleic Acids Research, 2003, Vol. 31, No. 13 **3659–3665**
DOI: 10.1093/nar/gkg531

MAK, a computational tool kit for automated MITE analysis

Guojun Yang and Timothy C. Hall*

Institute of Developmental and Molecular Biology and Department of Biology, Texas A&M University, College Station, TX 77843-3155, USA

Received January 8, 2003; Revised January 30, 2003; Accepted March 21, 2003

ABSTRACT

Miniature inverted repeat transposable elements (MITEs) are ubiquitous and numerous in higher eukaryotic genomes. Analysis of MITE families is laborious and time consuming, especially when multiple MITE families are involved in the study. Based on the structural characteristics of MITEs and genetic principles for transposable elements (TEs), we have developed a computational tool kit named MITE analysis kit (MAK) to automate the processes (http://perl.idmb.tamu.edu/mak.htm). In addition to its ability to routinely retrieve family member sequences and to report the positions of these elements relative to the closest neighboring genes, MAK is a powerful tool for revealing anchor elements that link MITE families to known transposable element families. Implementation of the MAK is described, as are genetic principles and algorithms used in its derivation. Test runs of the programs for several MITE families yielded anchor sequences that retain TIRs and coding regions reminiscent of transposases. These anchor sequences are consistent with previously reported putative autonomous elements for these MITE families. Furthermore, analysis of two MITE families with no known links to any transposon family revealed two novel transposon families, namely Math and Kid, belonging to the IS5/Harbinger/PIF superfamily.

INTRODUCTION

Higher eukaryotic genomes are rich in transposable elements. Two distinct types of transposable elements have been identified in higher eukaryotes: type I elements (retrotransposable elements) use a copy–paste approach to transpose, yielding a large copy number; type II elements (DNA elements) use a cut–paste–repair approach to transpose. However, numerous families of highly repetitive (hundreds or thousands), short (100–500 bp), elements that do not seem to belong to either type of element have been reported in plants and animals over the past decade (1–27). Because these families typically bear terminal inverted repeats (TIRs) and have target site duplications (TSDs) in their flanking sequences, they were given a collective name of miniature inverted repeat transposable elements (MITEs). Since MITEs apparently do not encode proteins, perhaps because of their small size, their amplification requires the involvement of factors supplied *in trans*. Although several MITE families are thought to be related to ancestral elements that bear similar TIRs and subterminal regions and are (or were) capable of coding for transposase-like proteins (20,23,24,28), the majority of MITE families lack such links. Since different MITE families may be derived from different founder elements, a link to the ancestral element needs to be established for each individual MITE family.

The analysis of a MITE family usually involves retrieving and aligning members in a given family, searching for its origin (or putative ancestor element) and studying its association with genes in a genome. These analysis steps are laborious, especially when multiple MITE families are involved in the analysis. To retrieve members of a family from the databases or to check the association of members with genes in a genome, a BLASTN is usually carried out, each high scoring pair (HSP) of the BLAST results is manually checked, and the desired sequence or positional information is then extracted from various accessions. This process is time-consuming and error prone because: (i) the copy number for MITE families is usually large; (ii) for purposes of alignment it is necessary to reverse the sequences of those hits that are on the complementary strand of the sequence; and (iii) for unfinished genomes, cited positions of elements in high-throughput sequences are subject to change until the genome is completely sequenced. Even for the announced genomes, updates are released frequently and, hence, the copy number, positions and annotation is subject to change. To search for the putative anchor element (that retains both TIRs and coding regions reminiscent of a transposase) for a MITE family *in silico*, a BLASTN is carried out and long elements containing similarity to both ends of the MITE are checked and are then used to do BLASTX. BLASTX is used to screen for similarity to known transposases. This process is also time-consuming because, in addition to the difficulties mentioned above, complications arise from the facts that: (i) the anchor element

*To whom correspondence should be addressed. Tel: +1 9798457750; Fax: +1 9798624098; Email: tim@idmb.tamu.edu

Table 1. Summary of information about MITE families used in this study

Family	Organism	Anchor element	Related transposase	Reference
Emigrant/MathE2	*A.thaliana*	AC006161 (85200-87313)	*Pogo*	(6,11)
mPIF	*Z.mays*	AF412282 (1-3725)	*PIFa*	(24)
Tc8	*C.elegans*	AF040643 (24047-31614)	*IS5*	(20)
MDM-2	*O.sativa*	AP004320 (3568-9016)	*MURA*	(23)
MathE1	*A.thaliana*	Unknown	Unknown	(11)
Kiddo	*O.sativa*	Unknown	Unknown	(22)

usually does not share internal sequence similarity to the query MITE element, thus the identification of long elements requires manual inspection and recording of the short BLAST HSPs; (ii) BLASTX searches usually take longer than BLASTN searches; and (iii) long sequences dramatically delay results from BLASTX.

Here, we describe the MITE analysis kit (MAK), a collection of programs designed to automate MITE analysis (http://perl.idmb.tamu.edu/mak.htm). Given the sequence of a MITE element, MAK can retrieve and orient sequences of other members of the family, identify genes closest to the MITE elements, and can predict the anchor element for the MITE family. Using MAK, we have identified two novel TE families named Math and Kid and provided evidence that they belong to the recently identified (24,29) IS5/Harbinger/PIF superfamily.

MATERIALS AND METHODS

Programming language and modules

Practical extraction and report language (Perl) (30) was used to write the programs for MAK. Transformation of sequence formats was carried out with Bioperl modules Bio::Seq and Bio::SeqIO. The module Bio::Tools::Run::RemoteBLAST was used to do remote BLAST searches and the modules Bio::Search and Bio::SearchIO were used to parse the BLAST search results. Bio::DB::GenBank was used to retrieve MITE elements and their flanking sequences. The Bio::SeqFeature module (31) was used to identify genes closest to the MITEs. Common gateway interface (CGI) programming (32) was used to set up the MAK web-based query service (http://perl.idmb.tamu.edu/mak.htm).

Computing resources

Database searches were executed in the queuing system for BLAST (QBLAST) (33) at NCBI using a Uniform Resource Locator (URL) standardized application program interface (API) (http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html#blastq). MAK was tested extensively with a UNIX system on a 48-processor SGI Origin 3800 (k2) supercomputer at the Texas A&M University supercomputing facility (http://sc.tamu.edu). It was also tested using either Linux or Win32 systems on a PC with 2 GB RAM at the Texas A&M University Institute of Developmental and Molecular Biology (IDMB). Manual BLAST was carried out at the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/) to confirm the results from MAK. AlignX in the VNTI7 package (InforMax, Bethesda, MD) was used for the alignments of DNA and protein sequences.

Data sets

Two sets of MITE sequences were used for this study. The families in the first group have reported links to known transposons and were used to test MAK anchor element prediction function. This group includes the families Emigrant/MathE2 (6,11), Tc8 (20), mPIF (24) and MDM-2 (23). The families in the second group did not have any reported link to a known transposon family at the time our study was carried out. This group includes families MathE1 (11) and Kiddo. The sequence of the dataset is supplied as Supplementary Material 1 and the information about the MITE families is summarized in Table 1.

RESULTS

Genetic principles and program pipelines

Member retriever. MITEs in a family share DNA sequence similarities that are readily detectable using BLAST searches. To illustrate the TIR conservation and relationship among members, an alignment is needed. Sequences of the members can be retrieved from BLAST search results. Since a BLAST hit can be on the top or bottom strand, all of the hit sequences to be used for alignment need to be in the same orientation; therefore, in MAK, the hits on the bottom (minus) strand are reversed. Since sequences adjacent to the TSDs of MITEs are often of interest to researchers, the program was designed to allow the retrieval of flanking sequences. In the Member retriever program, BLASTN searches are initiated against NCBI 'nt' and 'htgs' databases using a given MITE sequence. The search results are automatically retrieved and the HSPs are parsed. If the query sequence part in an HSP is the full query MITE length, the hit sequence in the HSP is retrieved as a complete element. If the hit part is in an opposite orientation (minus strand), the reverse sequence of the hit part is retrieved. Then, flanking sequences of user defined length are retrieved (Fig. 1A). In addition, long elements that do not show strong similarity along the total length of the query but do at both terminal regions can also be retrieved using the Long element function (Fig. 1B).

Anchor. MITEs are likely to be derived from various autonomous or receptor transposons. Recently, MDM-2, mPIF,

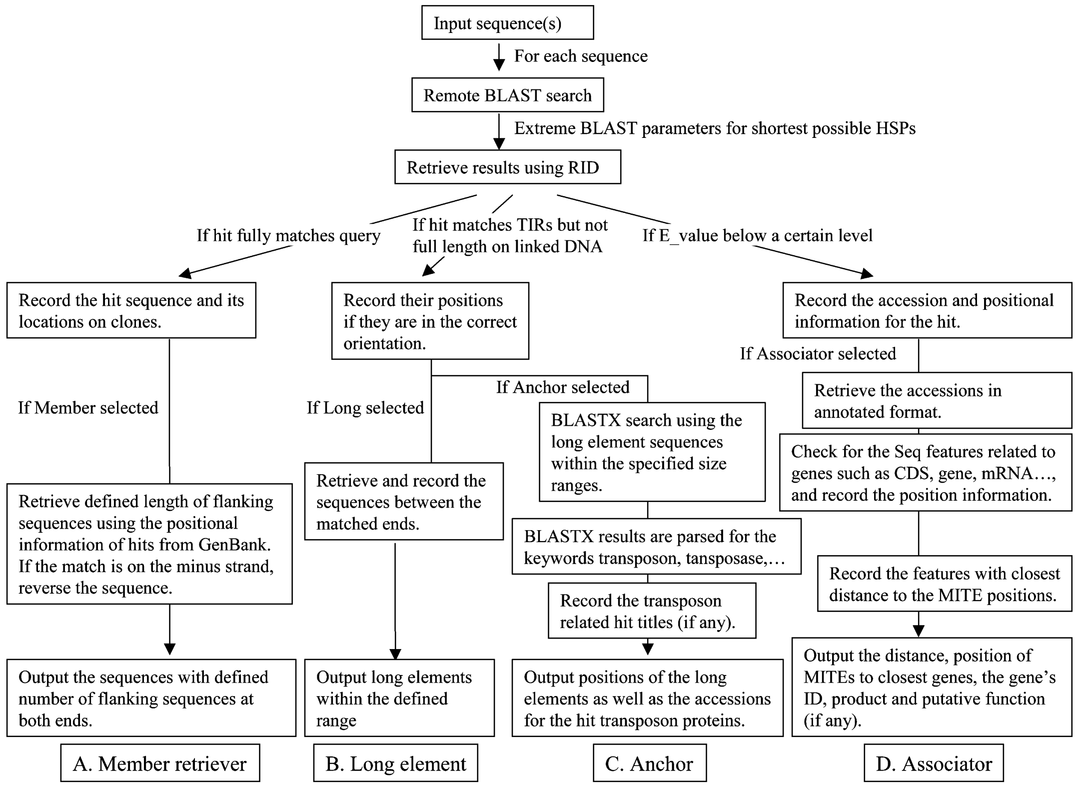

Figure 1. Diagram of pipelines for MAK.

Tc8 and Emigrant/MathE2 have been identified to be the derivatives of known transposons (20,23,24,28). The most conserved parts of a DNA transposon family lie in the TIRs because they represent the major transposase recognition sites. Since MITEs are usually so abbreviated that they retain no trace of the transposase coding region, identification of the original transposon relies heavily on their TIRs or subterminal regions. Since the elements from which MITEs are directly derived may not necessarily be the autonomous elements responsible for their transposition, we have denoted these elements as MITE anchors. While anchor elements may not necessarily be the ancestors of the anchored elements, an evolutionary relationship is likely to exist between the anchor elements and the anchored elements. In the automated anchor finding process, a BLASTN is carried out at very low stringency (with high E_value, lowest word size, high hitlist_size, low gapcosts and a specified organism) for shortest hits possible. All HSPs that match either end of the query sequence are checked for their orientation. If two HSPs matching the two ends of the query sequence in the proper orientation are identified from one accession number, the sequence between the HSPs (including the HSP regions) is retrieved if they are 100 bp longer than query sequence but do not exceed a total of the specified anchor size limit. A TIR tolerance is the maximal number of differences between the hits and the query outer ends. If a non-zero TIR tolerance is selected, matching to the query ends is less stringent. These long elements are possible ancestors or their 'uncles or cousins'. To determine if these elements have the potential to encode a transposase, a BLASTX is carried out for each of these elements and the hits that contain the word 'transposon', 'transposase' or 'transposable element' in their titles are retrieved (Fig. 1C). False predictions usually result from transposon nesting events and thus can be identified with BLASTN searches. If only the predicted transposase-like regions inside the predicted element are repetitive at the DNA sequence level, such entries are discarded. Since the long element retrieval process is based on matching with the terminal regions of the query, a nesting event involving two identical transposons with symmetric terminals will produce all the four possibilities if all of them are in the specified size range. Two of the desired elements can be easily identified through a BLASTN search.

Associator. Because of their short size, MITE families are potentially less disruptive than classic transposons and they may even contribute beneficially to gene regulation (34). Nevertheless, their large copy numbers suggest that they are potentially disruptive and very few MITEs have been found in coding sequences. It is often desirable to know how closely members in a family are associated with genes and which genes have closest proximity to MITE elements. In Associator (Fig. 1D), a BLASTN is carried out and the accessions and positions of significant hits with lengths longer than one fourth of the query are recorded. For each of these significant hits, the name and position of the annotated gene that has the closest proximity to the center of a given MITE element is retrieved. The results for all the significant hits can be exported as a table.

Implementation of MAK

MAK runs on UNIX, Linux and Win32 platforms on which Perl 5.6.1 (http://www.perl.com) and Bioperl 1.0.2 releases (http://www.bioperl.org) are installed. The web-based software (http://perl.idmb.tamu.edu/mak.htm) starts with the input of user name, email address, sequence file name and sequence(s). Then the desired function (Member retriever, Long elements, Associator or Anchor) needs to be selected. The parameters to run Member retriever include the length of sequence flanking the MITE members, the organism in which the MITE family is present and the terminal inverted repeat (TIR) tolerance. For Long element and Anchor functions, a size limit can be selected from 2000, 5000, 10 000 or 20 000 bp. All the retrieved long elements are at least 100 bp longer than the query sequences. Chosen E_value and organism parameters apply for all MAK functions. Upon initiation of the program, the user will be notified of the status of the process. The results will be sent to the specified email. While the format for the input sequence is flexible if the analysis is for a single MITE family, FASTA format is highly recommended if the analysis involves multiple MITE families. The MAK program can also be run as a queue job on a supercomputer in which multiple functions of MAK can be used to analyze several MITE families simultaneously.

When the dataset for MITEs (Table 1) was used to run MAK, updated information for these MITE families was obtained. The retrieved members of these MITE families are given as Supplementary Material 2. The chart in Figure 2 demonstrates the distance of MITEs (in completed genomes) from MathE1, MathE2 and Tc8 relative to their closest genes. The output from MAK Associator is provided as Supplementary Material 3. When the dataset for MITEs with known relationships was used to run Anchor function, anchor elements for Emigrant/MathE2, Tc8, mPIF, MDM2 predicted by the MAK were consistent with previous reports (20,23,24,28) (Table 1 and Supplementary Material 4).

Anchoring MathE1 and Kiddo MITE families

When the MathE1 family was used to run the Anchor function of MAK in the *Arabidopsis* genome, two identical long elements (AC007123, from 6918 to 2690; AF007271, from 16 996 to 21 224) with identical TIRs to the MathE1 element on accession AB010073 were identified. They were predicted to be a transposase gene. The sequence of these two long elements comes from an overlap region of accessions AC007123 and AF007271 on chromosome 5. Thus they represent only one element, which we named as A-MathE1 (anchor of MathE1). It shares 77% sequence identity to one terminus of MathE1 in 35 bp and 80% sequence identity to the other terminus of MathE1 in 82 bp. It has identical 12 bp TIRs to those of MathE1 elements. Interestingly, the internal sequences of MathE1 elements seem to be derived from 29 blocks of 10–30 bp on AC007123 with very little divergence (>90% identity in each block). BLAST searches with the long element sequences resulted in truncated or disarmed elements

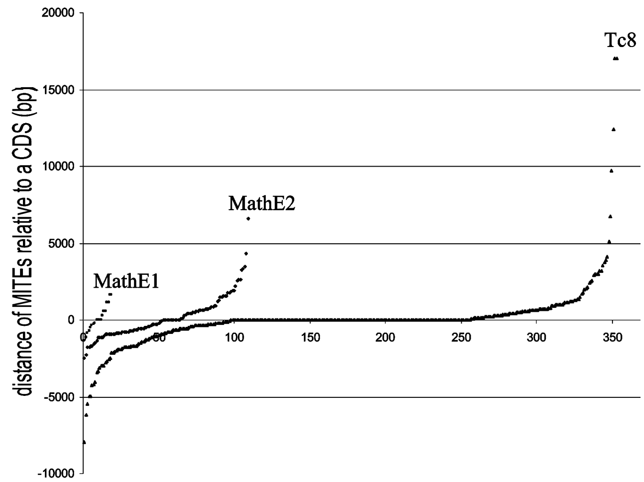

Figure 2. Distance of members in MITE families MathE1, MathE2 and Tc8 to their closest genes. The names and positions of the genes closest to the MITEs retrieved with MAK are sorted ascendingly with Microsoft Excel. The distance of a MITE inside a coding sequence (CDS) to the gene is considered 0 and the distance of a MITE at the 5′ end of a CDS to the CDS is changed into a negative value. The sorted elements are numbered consecutively, starting from 1. The distance values of MITEs to a CDS are plotted against their numbering. Each unit of *x*-axis on the chart represents a MITE element and the distance of that MITE to a CDS is shown as the value on the *y*-axis.

with minimal damages to the TIRs. An additional long element (AB025602, from 7658 to 11 849) showed an overall 98% DNA sequence identity to A-MathE1. Since the element on AB025602 is situated on a different locus of chromosome 5 from A-MathE1 on AC007123 and they share no flanking sequence similarity, it apparently results from a transposition event. The 6 bp missing from the TIR at the 5′ end were found to be present on the 3′ end flanking sequence. These long elements and MITE family MathE1 converge into one transposable element family we have named Math (Fig. 3A). This family showed a TSD exclusively of 'TTA' and has a TIR of 13 bp.

When the Kiddo family (22) was used to run the Anchor function of MAK in the rice genome, three long elements with typical TE characteristics were predicted to be transposase genes or pseudogenes. They are within AP004087 (gi:15281366) from 74 902 to 78 476 on chromosome 2, AC118347 (gi:20153328) from 20 719 to 17 088 on chromosome 11 and AP005461 (gi:21624013) from 78 912 to 82 646 on chromosome 6. These elements showed an overall sequence identity of >92%. When the long element of Kiddo on AP004087 was used to do a BLAST search, an additional complete element was found on AF114171 (gi:4680196) from 39 268 to 43 050 from *Sorghum bicolor* chromosome F. It has an overall sequence identity of 66% to that of AC118347. We name the long element on AC118347 as A-Kiddo (anchor of Kiddo). The internal DNA sequences of these four long elements are highly conserved in two regions (from ∼800 to ∼2050 bp and from ∼2100 to ∼3200 bp on AC118347; Fig. 3B). Additionally, 16 complete (i.e. having TIRs at both ends) elements with sizes ranging from 714 to 2538 bp were identified. Like the long elements, they have a consensus TSD of TAA and show high (>85%) similarity in ∼250 and

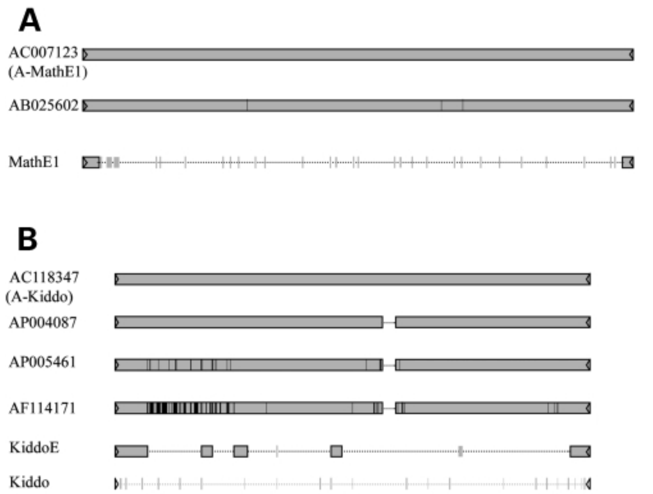

Figure 3. Schematic presentation of TE family Math (**A**) and Kid (**B**). Anchor elements are aligned with similar long elements and corresponding MITE families. Vertical lines in internal regions of long elements (long gray bars) indicate dissimilar regions and vertical lines connected by dotted lines in MITE elements indicate similar sequence blocks on MITEs to the anchor elements. Dotted lines indicate deletion regions (blank regions). The elements are drawn to scale. The triangles at the ends represent TIRs. The accession number on the left of the elements indicate the accession on which the elements are located and the positions for these elements on the accessions are described in Results.

~110 bp terminal regions, but their internal sequences do not show similarity to known transposases. They form an intermediate group (KiddoE) between Kiddo and A-Kiddo. Together, these elements represent a novel transposable element family (Fig. 3B), which we named Kid. These have not previously been annotated in the rice genome.

Math and Kid belong to IS5/Harbinger/PIF superfamily

The anchor elements of A-MathE1 and A-KiddoE do not share significant DNA sequence similarity with each other. Their internal sequences do not contain repetitive sequences as revealed by BLASTN searches. However, as predicted by MAK, both of them share strong similarity to putative transposase-like proteins. One of the BLASTX hits was from the putative transposase for maize PIFa elements (AF412282). The predicted gene structure for A-MathE1 and A-KiddoE is shown in Figure 4A and their putative translated products were aligned with putative PIFa protein, as shown in Figure 4B. PIFa shares a similarity of 46 and 50% in a region of 834 bp (808–1642) and 912 bp (2044–2956) to the putative A-MathE1 and A-KiddoE proteins, respectively. In addition, they have the same TSD size, and the TIR sequences of these two elements are very similar to those of PIFa (Fig. 4C). Indeed, the A-MathE1 element on AF007271 was proposed to be a IS5/Harbinger/PIF member named *At*-PIF2 (24). These pieces of evidence strongly suggest that the TE families Math and Kid belong to the IS5/Harbinger/PIF superfamily.

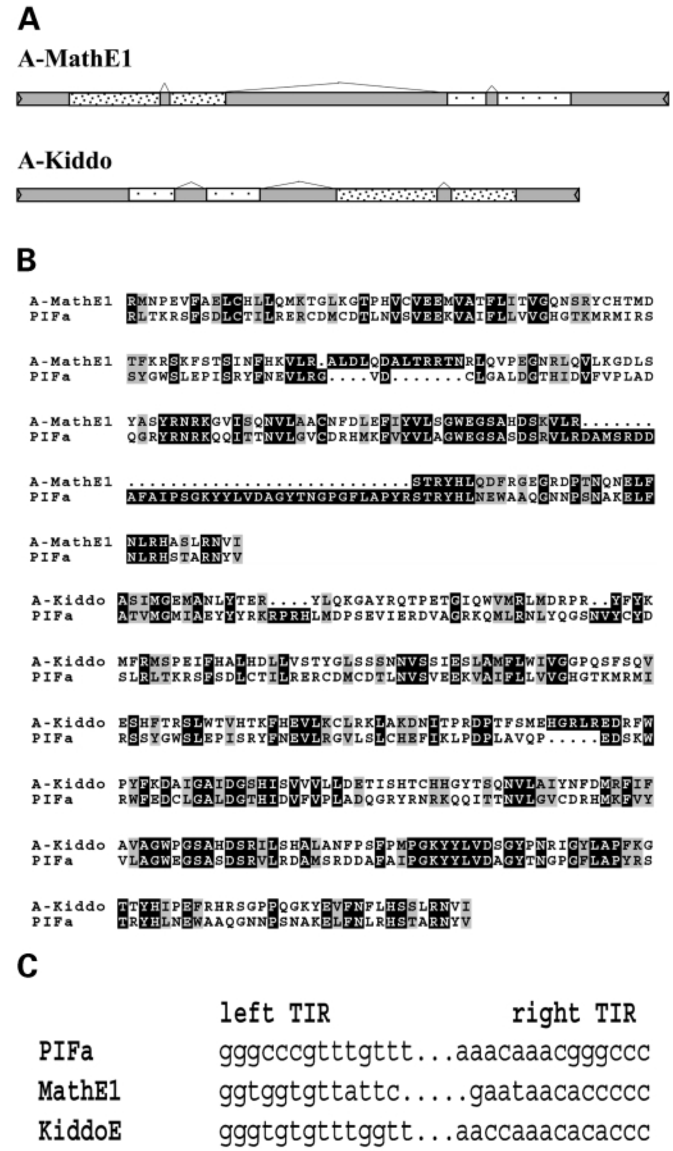

Figure 4. Putative gene structure for A-MathE1 and A-Kiddo (**A**). Dotted regions indicate putative coding exons. Exons showing similarity to putative PIFa transposase are indicated in densely dotted regions. Bridged regions indicate putative introns. Sequence alignment between A-MathE1 putative translation product (from 808–1642 on the DNA sequence) and maize PIFa putative transposase (from 70 to 296 on AF412282 protein sequence) (upper panel), and sequence alignment between A-Kiddo putative translation product (2044–2956 on the DNA sequence) and PIFa putative transposase (from 18 to 296 on AF412282) (**B**). Letters in black indicate identical residues and letters in gray indicate similar residues. Alignment of left TIRs and right TIRs from PIFa, MathE1 and KiddoE (**C**). Dotted lines denote omitted internal sequences.

DISCUSSION

Advantages and limitations of the computing approach

Using the automated processes in the MAK, we have successfully run a set of MITE families overnight. The output files are in standard format (e.g. FASTA) and thus can be used directly for downstream processes such as alignments (using

VNTI, PILEUP, ClustalW, etc.) and making tables. As noted in the Introduction, conventional MITE analysis is laborious and needs to be repeated each time the database is updated. Clearly, new analyses are appropriate as databases are updated, but this is relatively easy using MAK. However, since the process is partly dependent on remote BLAST analysis on the NCBI QBLAST server, the program may encounter internal server errors and hence be terminated (if this occurs, an error message will be generated). To lower the chance of encountering an internal server error at NCBI, we usually avoid running the program at peak times (usually daytime on workdays). In the program, we allow the retrieval of request ID (RID) for five times with an interval of 100 s before the process is allowed to die. Another alternative is to run stand-alone BLAST on a local system, but this approach requires downloading a huge database from NCBI. For the Associator function output, modest manual inspection to remove duplicate entries is necessary because BLAST searches will yield two HSPs at the same DNA locus if the MITE has a typical inverted repeat structure. Further improvement of the program to remove such entries is underway.

Misannotation of PIF-like elements in GenBank

When we used A-MathE1 and A-Kiddo elements to do BLASTX, several hits were titled *En/spm*-like transposon protein (accessions: NM_148036, NM_104832, NM_128220, NM_148535, AP003450, AB016878, AP000606, NM_148229). These *En/spm*-like hits were further analyzed using PSI-BLAST and iterations were carried out until no more new hits were found. Unfortunately, we were unable to find detectable peptide sequence similarity between any of these *En/spm*-like transposon proteins and the putative *En/spm* proteins TNPD-TNPA in maize (35), putative *Tam*1 proteins TNP1–TNP2 in *Antirrhinum majus* (36,37) or the putative open reading frame of *Tgm* in soybean (38,39). On the contrary, all of these hits showed strong similarity to PIFa putative transposase protein (AF412282). Together with the fact that A-MathE1 and A-Kiddo showed similar TIR sequence to that of the *PIFa*. We believe that these elements were misannotated in the database although it is still possible that *PIFa* and *En/spm* superfamiles are remotely related because they both have 3 bp TSDs and ~13 bp TIRs.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Charlie Harris at IDMB and David Reed at the Biology Department of Texas A&M University, and Hao Yu and Hemanth Sundaram at the Supercomputing Facility at Texas A&M University for exceptional help with computing techniques. We also thank Jason Stajich and the biperl community for advice on bioperl modules and Tom Bureau for help in obtaining the Tc8 sequences. Supported in part by NSF grant MCB-0110477.

REFERENCES

1. Bureau,T.E. and Wessler,S.R. (1992) *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, **4**, 1283–1294.
2. Bureau,T.E. and Wessler,S.R. (1994) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, **6**, 907–916.
3. Oosumi,T., Garlick,B. and Belknap,W.R. (1995) Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **92**, 8886–8890.
4. Bureau,T.E., Ronald,P.C. and Wessler,S.R. (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl Acad. Sci. USA*, **93**, 8524–8529.
5. Tu,Z. (1997) Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl Acad. Sci. USA*, **94**, 7475–7480.
6. Casacuberta,E., Casacuberta,J.M., Puigdomenech,P. and Monfort,A. (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. *Plant J.*, **16**, 79–85.
7. Song,W.Y., Pi,L.Y., Bureau,T.E. and Ronald,P.C. (1998) Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the *Xa21* family of disease resistance genes in rice. *Mol. Gen. Genet.*, **258**, 449–456.
8. Charrier,B., Foucher,F., Kondorosi,E., d'Aubenton-Carafa,Y., Thermes,C., Kondorosi,A. and Ratet,P. (1999) *Bigfoot:* a new family of MITE elements characterized from the *Medicago* genus. *Plant J.*, **18**, 431–441.
9. Iwamoto,M., Nagashima,H., Nagamine,T., Higo,H. and Higo,K. (1999) A *tourist* element in the 5′-flanking region of the catalase gene *CatA* reveals evolutionary relationships among *Oryza* species with various genome types. *Mol. Gen. Genet.*, **262**, 493–500.
10. Izsvak,Z., Ivics,Z., Shimoda,N., Mohn,D., Okamoto,H. and Hackett,P.B. (1999) Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.*, **48**, 13–21.
11. Surzycki,S.A. and Belknap,W.R. (1999) Characterization of repetitive DNA elements in *Arabidopsis*. *J. Mol. Evol.*, **48**, 684–691.
12. Tikhonov,A.P., SanMiguel,P.J., Nakajima,Y., Gorenstein,N.M., Bennetzen,J.L. and Avramova,Z. (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl Acad. Sci. USA*, **96**, 7409–7414.
13. Casa,A.M., Brouwer,C., Nagel,A., Wang,L., Zhang,Q., Kresovich,S. and Wessler,S.R. (2000) Inaugural article: the MITE family *heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl Acad. Sci. USA*, **97**, 10083–10089.
14. Elrouby,N. and Bureau,T.E. (2000) Molecular characterization of the *Abp*1 5′-flanking region in maize and the teosintes. *Plant Physiol.*, **124**, 369–377.
15. Feschotte,C. and Mouches,C. (2000) Recent amplification of miniature inverted-repeat transposable elements in the vector mosquito *Culex pipiens*: characterization of the *Mimo* family. *Gene*, **250**, 109–116.
16. Surzycki,S.A. and Belknap,W.R. (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl Acad. Sci. USA*, **97**, 245–249.
17. Zhang,Q., Arbuckle,J. and Wessler,S.R. (2000) Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions of maize. *Proc. Natl Acad. Sci. USA*, **97**, 1160–1165.
18. Akagi,H., Yokozeki,Y., Inagaki,A., Mori,K. and Fujimura,T. (2001) *Micron*, a microsatellite-targeting transposable element in the rice genome. *Mol. Genet. Genomics*, **266**, 471–480.
19. Jiang,N. and Wessler,S.R. (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell*, **13**, 2553–2564.
20. Le,Q.H., Turcotte,K. and Bureau,T. (2001) Tc8, a Tourist-like transposon in *Caenorhabditis elegans*. *Genetics*, **158**, 1081–1088.
21. Tu,Z. and Orphanidis,S.P. (2001) *Microuli*, a family of miniature subterminal inverted-repeat transposable elements (MSITEs): transposition without terminal inverted repeats. *Mol. Biol. Evol.*, **18**, 893–895.

22. Yang,G., Dong,J., Chandrasekharan,M.B. and Hall,T.C. (2001) *Kiddo*, a new transposable element family closely associated with rice genes. *Mol. Genet. Genomics*, **266**, 417–424.
23. Yang,G. and Hall,T.C. (2003) *MDM*-1 and *MDM*-2, two *Mutator*-derived MITE families in rice. *J. Mol. Evol.*, **56**, 255–264.
24. Zhang,X., Feschotte,C., Zhang,Q., Jiang,N., Eggleston,W.B. and Wessler,S.R. (2001) P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc. Natl Acad. Sci. USA*, **98**, 12572–12577.
25. Turcotte,K., Srinivasan,S. and Bureau,T. (2001) Survey of transposable elements from rice genomic sequences. *Plant J.*, **25**, 169–179.
26. Tu,Z. (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc. Natl Acad. Sci. USA*, **98**, 1699–1704.
27. Braquart,C., Royer,V. and Bouhin,H. (1999) DEC: a new miniature inverted-repeat transposable element from the genome of the beetle *Tenebrio molitor*. *Insect. Mol. Biol.*, **8**, 571–574.
28. Feschotte,C. and Mouches,C. (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. *Mol. Biol. Evol.*, **17**, 730–737.
29. Jurka,J. and Kapitonov,V.V. (2001) PIFs meet Tourists and Harbingers: a superfamily reunion. *Proc. Natl Acad. Sci. USA*, **98**, 12315–12316.
30. Schwartz,R.L. and Christianson,T. (1997) *Learning Perl*, O'Reilly, Sebastopol, CA.
31. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
32. Gundavaram,S. (1996) *CGI Programming on the World Wide Web.* O'Reilly, Sebastopol, CA.
33. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
34. El Amrani,A., Marie,L., Ainouche,A., Nicolas,J. and Couee,I. (2002) Genome-wide distribution and potential regulatory functions of AtATE, a novel family of miniature inverted-repeat transposable elements in *Arabidopsis thaliana*. *Mol. Genet. Genomics*, **267**, 459–471.
35. Pereira,A., Cuypers,H., Gierl,A., Schwarz-Sommer,Z.S. and Saedler,H. (1986) Molecular analysis of the En/Spm transposable element system of *Zea mays*. *EMBO J.*, **5**, 835–841.
36. Nacken,W.K., Piotrowiak,R., Saedler,H. and Sommer,H. (1991) The transposable element Tam1 from Antirrhinum majus shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol. Gen. Genet.*, **228**, 201–208.
37. Bonas,U., Sommer,H. and Saedler,H. (1984) The 17-kb *Tam1* element of *Antirrhinum majus* induces a 3-bp duplication upon integration into the chalcone synthase gene. *EMBO J.*, **3**, 1015–1019.
38. Rhodes,P. and Vodkin,L. (1985) Highly structured sequence homology between an insertion element and the gene inwhich it resides. *Proc. Natl Acad. Sci. USA*, **82**, 493–497.
39. Rhodes,P. and Vodkin,L. (1988) Organization of the *Tgm* family of transposable elements in soybean. *Genetics*, **120**, 597–604.